Trends in Genetics

CellPress
REVIEWS

Opinion

Replicability and Prediction: Lessons and Challenges from GWAS

Urko M. Marigorta,[1,8] Juan Antonio Rodríguez,[2,3,4,8,@] Greg Gibson,[1] and Arcadi Navarro[2,5,6,7,*]

Since the publication of the Wellcome Trust Case Control Consortium (WTCCC) landmark study a decade ago, genome-wide association studies (GWAS) have led to the discovery of thousands of risk variants involved in disease etiology. This success story has two angles that are often overlooked. First, GWAS findings are highly replicable. This is an unprecedented phenomenon in complex trait genetics, and indeed in many areas of science, which in past decades have been plagued by false positives. At a time of increasing concerns about the lack of reproducibility, we examine the biological and methodological reasons that account for the replicability of GWAS and identify the challenges ahead. In contrast to the exemplary success of disease gene discovery, at present GWAS findings are not useful for predicting phenotypes. We close with an overview of the prospects for individualized prediction of disease risk and its foreseeable impact in clinical practice.

Highlights

The decade of GWAS constitutes a clear improvement in the recent history of reproducibility in medical genetic research. The extremely high rates of replication imply that, for the first time, findings can be trusted.

Large numbers of false positives and the tiny effect size of genetic risk variants induced a change in incentives during the GWAS era, priming the requirement for large sample sizes and a culture of data sharing.

There is increasing interest in and need of new methodologies to better understand the genetic architecture of complex traits.

It is necessary to keep fostering a culture of compulsory replication to maintain the current high reliability in findings.

Although the success of GWAS has not translated into an ability to predict phenotypes based on genetic markers, polygenic and transcriptional risk scores (PRSs and TRSs) hold potential for stratification according to risk.

The High Replicability of GWAS Findings Is an Unprecedented Phenomenon

GWAS (see Glossary) are the tool of choice to identify genetic variants associated with complex disease and other phenotypes of interest [1]. They have revolutionized human genetics with the discovery of thousands of alleles that influence disease risk (Box 1) [1]. Success, however, has been accompanied by boom-and-bust mood swings in the field. The early crisis of the 'missing heritability' is a paradigmatic example [2]. The initial hype after the publication of the WTCCC [3] was dampened by generalized disappointment about the amount of **genetic variance** explained by GWAS results, which was perceived as disappointingly low. Despite the exponential yield in terms of discovered loci, the relevance of GWAS findings often remains under the spotlight [4].

There is, however, a persistent feature of GWAS results that is often overlooked: they stand out for their high **replicability**. The past decade has been characterized by a plethora of findings that have stood the test of time almost in their entirety. This degree of success is an unprecedented phenomenon in the field of complex trait genetics and, even more generally, in many scientific fields [5], which some argue are plagued by false-positive findings that generate a sense of 'replicability crisis' that may be fuel for antiscience movements (Box 2) [6].

Here we examine the biological and methodological reasons that have led to the success of GWAS in replicating genetic discoveries. We start with an analysis of the replicability of disease-associated variants discovered by GWAS, with special attention to temporal and spatial patterns of replication and the inferences that can be drawn about the **genetic architecture of disease**. We next discuss the reasons behind the outstanding reliability of GWAS findings and identify extant challenges to ensure the replicability of future findings. We end with an analysis of the current state of phenotype prediction based on GWAS findings, which remains one of the key challenges for successful implementation of precision medicine.

[1]Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, GA, USA
[2]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Catalonia, Spain
[3]Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain
[4]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
[5]Institute of Evolutionary Biology (UPF-CSIC), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

 https://doi.org/10.1016/j.tig.2018.03.005

Box 1. Quick Glance at the History and Rationale of GWAS

The two-page perspective by Risch and Merikangas in 1996 [59] can be considered as the start of the GWAS era. This landmark paper demonstrated that association studies based on polymorphisms spanning all of the genome would be powered to detect susceptibility variants of small effects. The key idea for association mapping lies in exploiting the correlation structure among markers that are nearby in our genomes (LD). This property allows the performance of indirect association tests, whereby it is necessary to type and test only a single variant (tagSNP) selected to act as a surrogate for all variants in the block of LD.

GWAS look for alleles transmitted with disease status and therefore with significantly different frequencies between cases and controls. After imputation of variants, GWAS test millions of SNPs and hence it is important to account for multiple testing to avoid false positives. The most used threshold to declare statistical evidence of association is $P < 5 \times 10^{-8}$, which arises from adjusting a 5% type I error rate (false positives) to the roughly 1–10 million independent tests performed.

Attention to other statistical matters is also important. Among others, extensive quality control steps include ensuring the reliability of the genotyping data, careful evaluation of the genetic ancestry of the samples to avoid population stratification, which may lead to *P*-value inflation if unaccounted for, and adjustment for clinical covariates such as age, gender, and lifestyle to avoid spurious associations and biases in the estimation of effect size. Finally, GWAS are increasingly performed through large meta-analyses that combine statistical evidence from multiple cohorts. Even if all substudies test the same hypothesis, it is important to quantify the presence of heterogeneity to identify outliers that should be excluded.

Overall, the GWAS era has constituted an impressive *tour de force*. As beautifully shown in the famous karyotype diagram released quarterly by the GWAS Catalog,[ii] we now know of over 12 000 SNPs associated at genome-wide significance with myriad traits. Overall, GWAS have demonstrated a sweeping performance in human genetics and are likely to remain as the main tool for the discovery of genes involved in disease in the future.

Widespread Replicability of GWAS Findings

Replication of newly reported results is the most reliable validation of scientific discoveries, as it confirms their true-positive status. In short, 'there is consensus in science that the final arbiter is replication' [7]. In complex trait genetics, the most used definition is exact replication, whereby the same genetic marker is consistently associated with the same phenotype in independent

Box 2. Replicability Crises in Other Scientific Fields

In 2010 Carney *et al*. reported that adopting an expansive body posture before facing a stressful situation could boost self-confidence and authority by increasing testosterone and decreasing cortisol levels [60]. Despite attracting plenty of attention, this finding failed to replicate in a larger study on power posing and hormonal levels [61]. Similar to social psychology, neuroimaging has also been on the spot in recent times. Of note, similar false positives have been ubiquitous across scientific fields [5], including physics research in the 1960s through the discovery of 'polywater', a hypothetical polymerized form of water that was eventually debunked [62].

Genomics itself has not escaped from the lack-of-replication problem, including the highly cited and yet still-debated gene-by-environment interaction between the serotonin transporter gene, adverse life events, and risk of depression [63]. In this regard, lack of correction for multiple testing is often the primary factor leading to false positives. In many fields, $P < 0.05$ is still deemed enough evidence to call an association true. Given that studies are generally under-powered, using such a liberal threshold is a recipe for false positives [64]. Genomics and physics have shielded themselves against this problem by imposing highly stringent *P*-value cut offs, setting an example that should be embraced by scientists in other communities [65].

Some scholars have stated that this replicability crisis is overblown [66]. A major issue is that controlling the false positive rate (FPR) depends on the fraction of hypotheses tested that are true [5]. Thus, if only 1% of hypotheses prove to be correct, with 80% power at $P < 0.05$ only eight of 1000 tests will be true positives, relative to 50 false positives. Happily, statistical analysis of reported *P*-value distributions strongly implies that the true-positive rate is much higher and the FPR is well controlled across much of science [67]. Also on the bright side, the consensus about the importance of ensuring large enough sample sizes is steadily growing. In this regard, several initiatives to detect false positives have emerged, namely the Reproducibility Project, Pubpeer, clinicaltrials.gov, and Retraction Watch, among many others. Taking all of this into account, if we want to assure a future of sustainable, reproducible, and replicable science we need to take direct action and extend the basic principles that have been shown to work to as many fields of research as possible.

[6]National Institute for Bioinformatics (INB), Barcelona, Catalonia, Spain
[7]Institució Catalana de Recerca i Estudis Avançats (ICREA), PRBB, Barcelona, Catalonia, Spain
[8]These authors contributed equally
@Twitter: @jrotwitguez

*Correspondence:
arcadi.navarro@upf.edu (A. Navarro).

cohorts [8]. This is usually attempted in GWAS, which usually have built-in replication stages for markers with the strongest evidence of association. However, these follow-up attempts do not necessarily adhere to the exact replication criterion and less stringent significance criteria are often used. Often, as well, consistency of the direction of effect can lead to an initially marginal association attaining **genome-wide significance**, which is widely regarded as replication.

Although external replication is always advantageous, clear-cut evidence is often lacking because diverse definitions of replication are used in the literature. For instance, the statistical evidence warranted to declare replication may range from a soft threshold, whereby nominal association (e.g., $P < 0.05$) is deemed enough by subsequent studies, to a hard one in which only associations that repeatedly achieve genome-wide significance (and are hence 'rediscovered') in several studies are considered formally replicated. To complicate matters, positive association at variants in strong **linkage disequilibrium (LD)** with the original marker SNP is often used as evidence of replication even if formal exact replication is not achieved. In this Opinion article, we use the '**rediscovery**' label for whenever a previously discovered SNP–trait association achieves strong statistical significance in a subsequent study and always explain the criteria used to consider replication when discussing the inferences from studies analyzing GWAS replicability.

Due to these hard-to-reconcile criteria, the replicability of GWAS is an important dimension that often goes without further scrutiny. A simplistic but straightforward way to assess replication is to use the NHGRI-EBI GWAS Catalog[i] for evaluation of the degree to which GWAS hits are rediscovered in subsequent studies. In Figure 1 (Key Figure), we present a temporal analysis of the catalog, a widely used resource based on manual curation of SNP–trait findings from published GWAS. Since the catalog includes only those associations with suggestive statistical evidence of association ($P < 10^{-5}$), any record in the catalog that reports a SNP–trait pair that was reported in a previous study can be labeled as a positive replication with strong statistical evidence of the previously known association. The figure includes all findings for 60 diseases with >15 discovered SNPs. To allow for array heterogeneity, we label as replication any record where an SNP in LD with the discovery SNP ($r^2 > 0.8$) is associated with the same disease in a subsequent study. The figure illustrates that most of the top hits included in the GWAS Catalog had already been reported in previous publications and hence correspond to replications of known SNP–trait associations. This was the case even in the first years of GWAS, which were often met with skepticism because the yield of new associations was perceived as unsatisfactory [9].

Perhaps a more underappreciated aspect of that early period of the GWAS era is that, for the first time in complex trait genetics, findings are undoubtedly robust. For instance, all of the 24 risk associations discovered at genome-wide significance by the WTCCC in 2007 were replicated within 2 years. This pattern has not changed over the years, and the bulk of findings since 2012 are replications of associations that have been described at least twice.

Robustness is widespread and cannot be explained only because large-effect associations replicate extensively. An inspection of the replicability of six disease traits illustrates this point, which holds for rediscovery in the GWAS Catalog at $P < 10^{-5}$ as well as for nominal replications at $P < 0.05$ in large meta-analyses published after the discovery study. Specifically, we compared the average rediscovery rates of variants of moderate [odds ratio (OR) between 1.2 and 1.5] and low **effect size** (OR < 1.2). As shown in Figure 2, rediscovery rates in the GWAS Catalog are similar regardless of effect size; namely, an average of 36.5% across the six diseases for variants of moderate risk and 43% for those of weakest effect ($P = 0.93$, Kolmogorov–Smirnov test).

Glossary

Candidate-gene studies: the main strategy used before the GWAS era. These studies tested only a small set of mutations in a genic region with plausible involvement with disease.

Common disease–common variant (CD/CV): the model that disease is attributable to a moderate number of variants of common frequency. This was the main paradigm under the studies in the early 2000s and it played a key role in fuelling the genome-wide strategy, although it is no longer considered relevant.

Effect size: a quantitative measure of the strength of a phenomenon. In GWAS it has been generally reported as OR for binary traits (diseases) and as regression coefficient (β) for quantitative traits.

Genetic architecture of disease: the underlying genetic basis of a phenotypic trait, which includes the total number of causal variants, their frequencies and magnitudes of effect, and their degree of interactions with one another or environmental factors.

Genetic variance: the proportion of total variance in a certain phenotype that is explained by the combination factors.

Genome-wide association studies (GWAS): analyze hundreds of thousands of markers, generally SNPs, comparing their frequencies across very large samples of individuals that share phenotypic characteristics.

Genome-wide significance: a P value that is generally used as a threshold to consider a positive finding. In GWAS it is usually declared when a variant is associated with $P < 5 \times 10^{-8}$. This value is considered robust to avoid multiple testing issues and is roughly the Bonferroni value for 0.05 divided by ~1 million independent genomic intervals.

Linkage disequilibrium (LD): the nonrandom association of alleles at different loci. The levels of LD vary among loci depending on local recombination rates. In humans, nearby genetic markers are usually linked in blocks of high LD, a feature that facilitates disease variant discovery through LD-based association mapping.

Rediscovery rates of ~40% would seem to imply that more than half of GWAS hits remain unreplicated. However, these percentages are severe underestimates because the GWAS Catalog records only those SNP–trait associations that achieve $P < 10^{-5}$. A more detailed analysis of the same traits focused on recently published meta-analyses renders much higher replication rates. In total, we enquired the status of 339 SNP–trait associations discovered at genome-wide significance ($P < 5 \times 10^{-8}$) by GWAS published before the corresponding meta-analysis study. An impressive 94% of results do indeed replicate at $P < 0.05$ (Figure 2 and Table S1 in the supplemental information online) with high correlation between the discovery and replication attempt ORs (Pearson's $r = 0.54$, $P < 10^{-16}$). These magnitudes resemble the ~100% replication rates observed in the literature after accounting for statistical power and the winner's curse [10,11], confirming the reliability of GWAS. The relatively low attention that replication receives together with the confusion in the literature about what can be formally considered as replication suggests that initiatives such as the GWAS Catalog should formally track replication of their SNP–trait entries.

Cross-Ancestry Replicability

Although more challenging, comparisons across ethnic groups reinforce the same idea and, furthermore, show that risk variants are shared across ancestries. Early analyses of the **Population Architecture using Genomics and Epidemiology (PAGE) multiethnic cohort** reported perfect concordance in risk allele directions at risk SNPs for prostate cancer and type 2 diabetes (T2D) [12,13]. This consistency holds for a wide range of diseases [14] and large nominal ($P < 0.05$) replication rates are observed at the level of SNPs [10], loci pinpointed by these SNPs [15], and genome-wide effect sizes [16]. A caveat, however, is that GWAS based on individuals of non-European ancestry usually study fewer individuals [17], rendering our picture of genetic susceptibility to disease across populations incomplete [18]. Moreover, caution is required since incomplete replication can also be informative: there are sound cases of lack of interpopulation replicability indicating that some risk variants are population specific (e.g., NOD2 and Crohn's disease in East Asian populations [19,20]). Correlations in effect size across ethnicities are often significantly less than one, and accordingly a fraction of SNPs do not replicate. For example, in the comparison of European and Asian associations with major depression, failure to replicate is most commonly because of divergent allele frequencies, which reduce power in one population since the proportion of attributable risk declines with minor allele frequency [21].

Replicability Is a Function of the Genetic Architecture of Disease

Besides confirming the reliability of GWAS as a tool to dissect genetic risk, patterns of temporal and cross-ancestry replicability can illuminate the genetic architecture of complex traits, informing about the reliability of effect estimations and their variability across human ancestries. Age-related macular degeneration (AMD) is a case in point. In one of the first GWAS, Klein *et al.* [22] detected a strong risk variant located in *CFH* (OR = 4.6). Given their exceedingly low sample size (96 patients and 50 controls), this report could have been a fluke or, worse, a severe overestimation of the real effect size. With time, however, this and a few other high-risk variants have been extensively replicated, indicating that a large proportion of AMD heritability is accounted for by a handful of variants with extremely large effect sizes [23].

More generally, the power to discover an association is proportional to the amount of phenotypic variance explained by the polymorphism; namely, the product of the effect size squared and the heterozygosity ($VE = 2pq\beta^2$). Here the effect size in standard deviation units is the average effect of substituting one allele for the other. For a trait regulated by thousands of genes, even for common alleles, β will generally be smaller than 1/20th of the standard

Population Architecture using Genomics and Epidemiology (PAGE) multiethnic cohort: the PAGE study is an ongoing consortium formed by several multiethnic population-based studies. The PAGE consortium has published several pioneer studies exploring the consistency of risk effects across genetic ancestries.

Publication bias: phenomenon in which only a fraction of studies performed is eventually published and available to other researchers, usually involving studies that fail to find an effect in the expected direction remaining unpublished.

Rediscovery: the finding of SNP–trait associations with suggestive statistical evidence or even genome-wide significance that is consistent with the discovery from the original study.

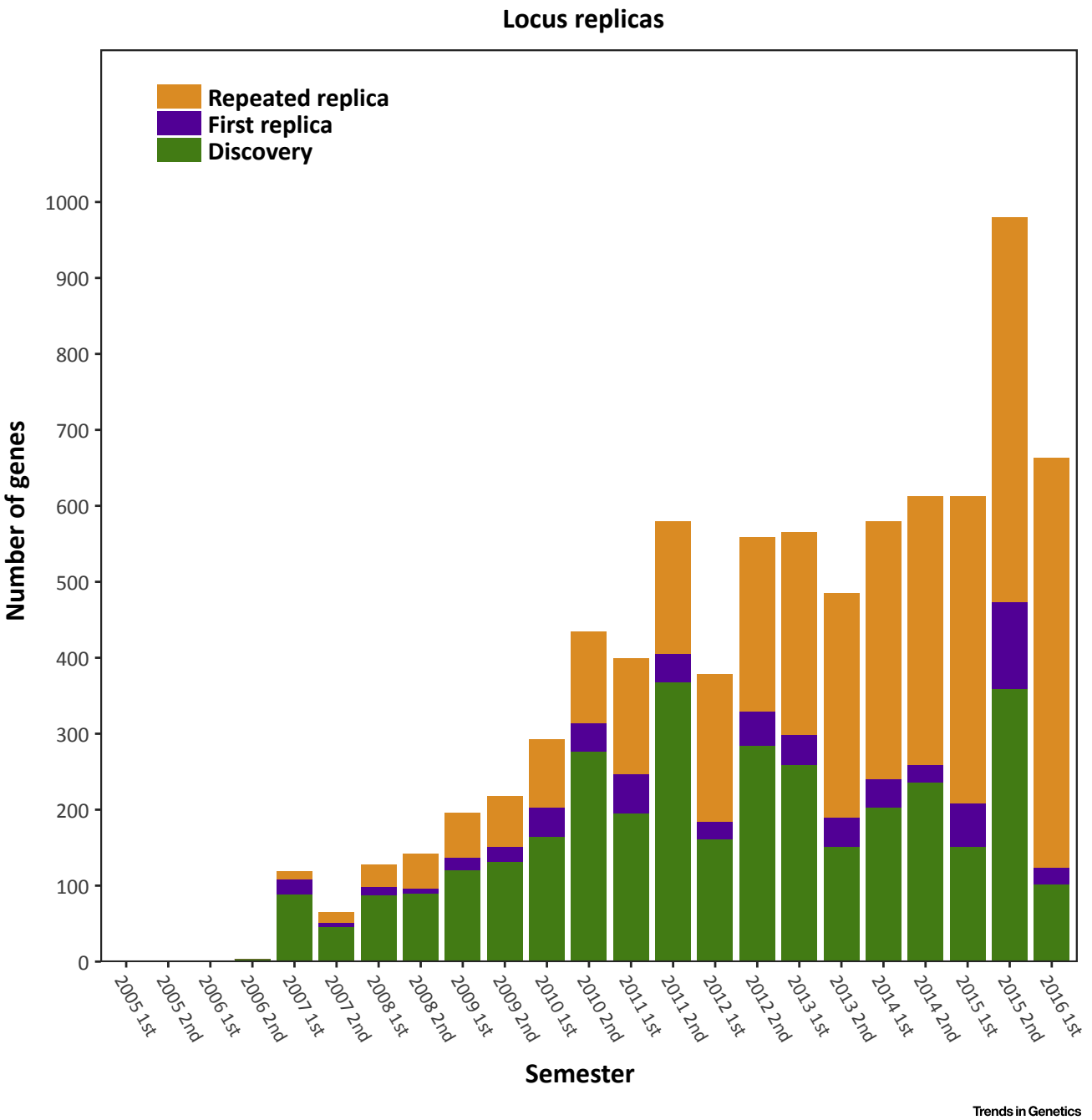
Replicability: property of a study or field whereby it produces the same or similar results when reproduced in a different sample, location, or time.

Transcriptome-wide association study (TWAS): a new methodology that uses genotypes and summary-statistic-level data to impute gene expression in individuals participating in GWAS. This permits the identification of significant gene–trait associations without the harsh multiple-testing correction typical of GWAS.

Type I error: statistical term that refers to the situation in hypothesis testing in which the null hypothesis is rejected despite being true (often called a false positive). Given the large number of independent tests performed, the probability of type I errors increases exponentially in GWAS if multiple-testing corrections are not considered.

Key Figure

An Increasing Proportion of Genome-wide Association Study (GWAS) Findings Corresponds to Replications of SNP–Trait Associations That Were Previously Reported



(See figure legend on the bottom of the next page.)

deviation (e.g., the equivalent of less than 1 mm of height). Very large sample sizes are required to detect such effects and there is usually a winner's curse of overestimating the effect size, but suitably large replication studies can confirm their influence. Replicability thus also informs about the shared architecture of other diseases vastly more complex than AMD. For instance, the convergence in associated loci confirms that a large proportion of the genetic risk for immune diseases is shared across conditions [24,25]. Similarly, the relative lack of findings for mental disorders suggests that psychiatric traits have an infinitesimal architecture controlled by many loci of tiny effects that can be discovered only through large meta-analyses [4,26].

A corollary is that patterns of replicability at specific associations can afford valuable lessons about disease heterogeneity. For instance, in 2007 three contemporaneous scans for T2D by the WTCCC, DGI, and FUSION consortia collaborated in one of the first coordinated data-sharing efforts to enhance evidence about the detected loci [3,27,28]. From the incomplete interstudy replication for the signal at the *FTO* locus, with clear heterogeneity in effect, it was learned that FTO exerts its effects through body mass index (BMI), playing only an indirect role in T2D risk [29]. More recently, the development of new methodologies to systematically explore intercohort and interdisease heterogeneity suggests that studying replicability will be key to improving our understanding of the genetic etiology of disease [30,31].

Collaborative Science and Large Cohorts Were Incentives for Reproducibility

The reliability of GWAS findings is a novelty that departs sharply from historical experience in the field. For over two decades, the literature in medical genetics has been tormented by promising but later discredited discoveries [32,33]. This was specially the case for **candidate-gene studies**, in which research groups would perform association analysis focused on genes that had been selected previously based on their biological plausibility. These studies in particular have been characterized by high rates of false positives [34]. For instance, after thousands of candidate-gene studies published during the 1990s and early 2000s, only a handful of the most-studied associations have shown consistent replication [32]. Moreover, the small fraction of findings from the candidate-gene era that have been confirmed by GWAS [35] mainly comprises risk variants of low effect sizes discovered by the studies with the largest sample sizes [36]. Many factors have contributed to the poor track record of candidate-gene reproducibility, including insufficient availability of genetic markers, inadequate handling of population structure, lack of statistical power due to low sample size, improper control of multiple testing, and extensive **publication bias** [32].

The surplus of promising but eventually failed associations seriously undermined the credibility of the whole LD–association-mapping approach, but on the bright side it made researchers aware that they needed to do better. Awareness about the ubiquity of false positives and about effects being smaller than initially expected shaped the field profoundly [pre-GWAS simulation

Figure 1. The graph reports the number of discoveries of new loci and rediscovery of previously discovered loci for 60 diseases included in the GWAS Catalog (quantitative traits were not considered; last accessed 17 March 2017). Data are classified according to the semester of publication (from 2005 to 2016). Given that all records included in the catalog achieve at least $P < 10^{-5}$, all newly included SNP–disease pairs that were already included because of being discovered by previous GWAS can be considered replications of the first finding. We labeled these instances as rediscoveries or replicas. The rediscovery figure for a given semester corresponds to the total cumulative number since 2005, separated according to whether the rediscovery event constitutes the first evidence for replication of a given locus ('first replica') or whether it was previously replicated ('repeated replica'). Given the diversity of arrays before the recent generalization of imputation in GWAS, SNPs are considered replicated when either the same SNP is rediscovered or another SNP in $R^2 \geq 0.8$ (using 1KG Europeans) within a ±500-kb window is reported. Only GWAS performed in Eurasian populations were evaluated.

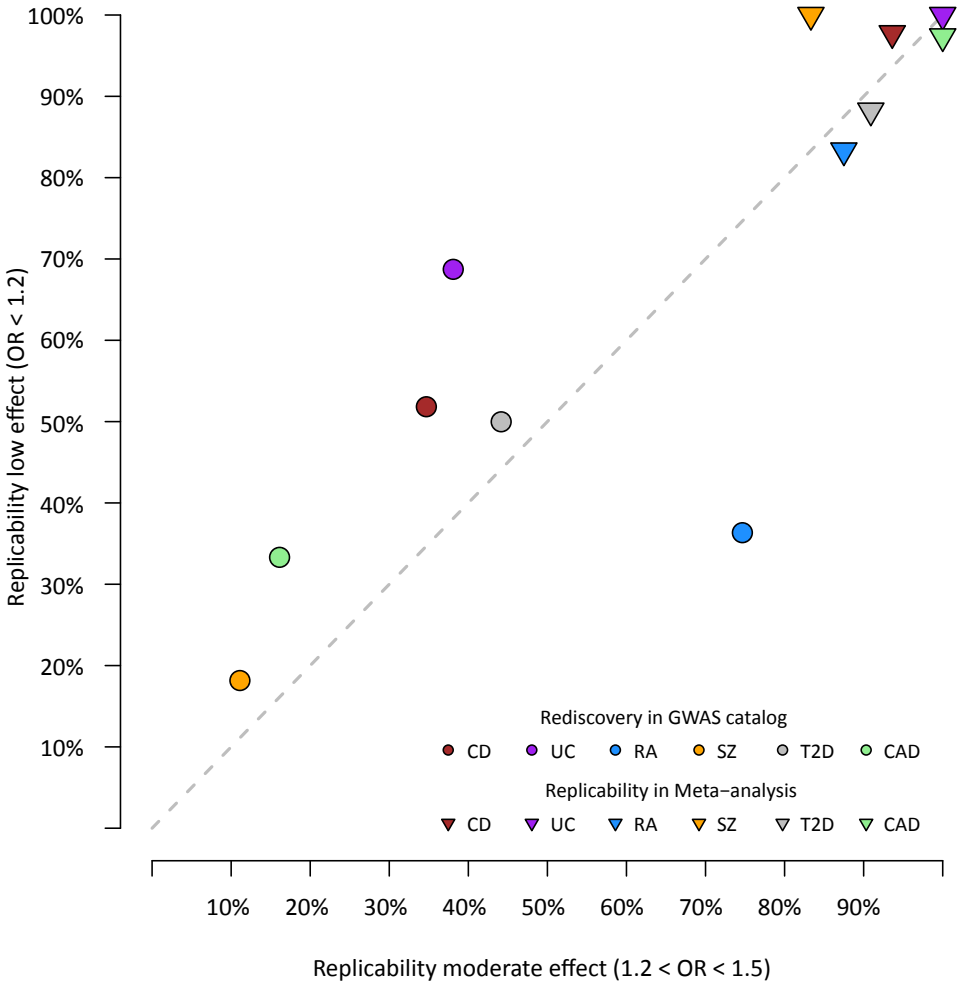


Figure 2. Risk Variants Discovered by Genome-wide Association Studies (GWAS) Replicate Similarly Regardless of the Odds Ratio (OR). X-axis: Classified by disease, percentage of SNPs with discovery OR between 1.2 and 1.5 that have been rediscovered at least once and included in the GWAS Catalog (filled circles) or that achieve nominal evidence of association ($P < 0.05$) in a large meta-analysis published after the discovery GWAS (inverted triangles). Y-axis: For the same diseases, proportion of rediscovery for risk SNPs with discovery OR < 1.2. Although the percentages vary by disease, the average rediscovery of low-effect variants is not significantly lower than that of variants with larger effect ($P = 0.93$, Kolmogorov–Smirnov test) and overall rediscovery estimates are highly correlated (Spearman's rho: 0.54, $P = 10^{-16}$).

studies concurrent with the development of the **common disease–common variant (CD/CV)** hypothesis systematically assumed ORs of 1.5, 2, 4, or larger] [37]. Such awareness led to: (i) a push for increased sample size; (ii) the use of stringent measures to control for multiple testing and avoid false positives (**type I error**); and (iii) a culture of collaboration and data sharing [36].

While it is also true that technological advances in genomics helped to pave the way for GWAS, the generalization of proper multiple testing corrections and setting adequate *P*-value thresholds to declare significance might have been the most important factor favoring the large replicability of GWAS. The push for statistical rigor was embedded in a new culture of data

sharing and demands for external replication before publication (often enforced by funders [38] and journals [39]), which not only translated into more powered studies but also reduced the incidence of publication bias. For instance, the strict threshold to declare genome-wide significance proposed in the original WTCCC study and accompanying feature [40] was a clear guideline against false positives adopted by subsequent studies.

As a final point, the expectation that genetic effects should for the most part be very small has allowed the flagging and correction of results that looked flawed, such as the original GWAS for exceptional longevity [39]. Collaboration and data sharing, once restricted to large-scale scientific infrastructures such as large particle accelerators, has more recently been spearheaded by the genomics community. We propose that similar changes would benefit other scientific areas. This is not to say that we simply favor 'big science', but rather that we believe all scientific endeavors, regardless of their scale, benefit from openness.

What Next? Replicability in the Full-Genome Era

The cost-effectiveness of SNP array genotyping followed by imputation suggests that GWAS will remain for some time as the main instrument for the detection of loci associated with complex traits [1,41]. New technological advances, however, may challenge the replicability achieved by the current generation of GWAS. For instance, the decreasing cost of high-throughput sequencing (NGS) allows exome and whole-genome studies that will, however, need new thresholds to correct for multiple testing [42]. In addition, the availability of ample clinical history for some big cohorts, such as the UK Biobank, asks for careful control of spurious correlations that may arise from multiple comparisons.

New methodological approaches can also potentially harm the reliability of findings. For instance, the interest in nonadditive effects such as epistasis and gene-by-environment interactions demands careful attention to latent covariates, which can introduce biases if unaccounted for [43,44]. Similarly, the list of new tools, such as multimarker analyses, which go beyond single SNP tests, or the inclusion of functional evidence to reweight GWAS results, is growing by the day [45,46]. The heterogeneity in these methodologies will necessarily complicate the evaluation of replicability.

Strict guidelines for publication and the ample experience gathered suggest that GWAS will maintain large replicability rates. However, exponential increases in available genomic and phenotype information may lead to progressive specializations in the hypotheses tested, necessarily departing from a classical GWAS framework. For instance, investigators working on exome data might decide to increase statistical power by aggregating genes into pathways. Under such scenarios, progressive relaxation of the statistical standards may lead to higher rates of false positives.

Extremely large GWAS finding hundreds of signals are already neglecting the importance of ascertaining the positive replication status of each discovered variant. For instance, the 2014 meta-analysis for height by the GIANT consortium focused on the overall concordance of effects between discovery and independent cohorts [47]. Instead of discussing the replication status of specific findings, this study highlighted the overall amount of phenotypic variance explained [47]. As explained in the next section, this approach is growing in parallel with the interest in using GWAS data for phenotype prediction. However, for variant and gene discovery purposes, insisting on strict significance thresholds and on garnering replication evidence for each individual variant will still be the best recipe to ensure high replicability rates.

The Problem of Low Predictability

As the biological underpinnings of disease are progressively unveiled, using the findings from GWAS to predict risk of disease at the individual level has gained traction. However, despite its importance for the prospects of implementing precision medicine in the clinic, phenotype predictability remains very low [48]. In the following paragraphs we summarize the reasons for the current state of phenotype prediction. We provide specific examples and calculations to visualize the performance of current genetic predictors and end with a discussion of the prospects for using predictors in the clinic.

The most-used method for risk profiling involves the calculation of genetic risk scores (GRSs) that sum the number of genome-wide significant alleles per individual in a target cohort. However, even if cases tend to have larger risk scores on average, the distributions of cases and controls tend to overlap for the most part. Although reasons related to the genetic architecture of complex traits partially account for the limited power of GRSs, this is also an unfortunate consequence of the limited fraction of disease susceptibility that GWAS have uncovered.

One of the most-used measures to test the performance of any risk score is the area under the ROC curve (AUC), which summarizes the true-positive rate (sensitivity) and false-positive rate (1 – specificity) for all possible cut offs. A classifier performing randomly and thus with null predictive power has an AUC of 0.5, whereas a perfect one reaches an AUC of 1. Thanks to the public availability of summary statistics from GWAS, it is straightforward to explore the performance of GRSs in any target dataset. In Figure 3 we use genotypes from the WTCCC to offer a glimpse of the temporal progress of GRSs and resulting improvements in AUC values for five different conditions. In unbroken lines, we show the evolution of classical predictors based on genome-wide significant SNPs. The plot indicates that, although the list of risk SNPs available in the GWAS Catalog has increased steadily over the years, the power of GRSs seems to have plateaued and only tiny improvements in the ability to distinguish between cases and controls can be noticed for the most recent years. In the broken lines, we add the equivalent calculations using variants detected by the COMBI algorithm that was recently proposed [45]. This method adds a preliminary step based on a support vector machine that aims to improve the detection of variants and that, as seen in Figure 3, has the potential to refine GRSs and provide gains in performance.

Given the low power of GWAS to detect small effects, most causal variants are known to remain as false negatives. In recent years, genetic prediction has moved towards increasing the signal by using markers that do not achieve genome-wide significance. In this regard, polygenic risk scores (PRSs) constitute a more powered alternative to GRS. PRS are calculated using SNPs selected through less-stringent significance thresholds (e.g., $P < 10^{-3}$). However, even for PRSs using thousands of SNPs, the AUC for most traits is still below 0.7 [49]. These low figures indicate that GWAS have not yet mustered the large sample sizes – in the millions – that are needed to refine effect-size estimates and achieve accurate individual risk profiles [50].

Although larger GWAS will lead to better characterization of the genetic architecture of disease, the field of genetic prediction is moving towards using evidence from genome-wide markers. A flurry of new methods is already using publicly available summary statistics from large studies [51], incorporating evidence from LD patterns to refine the effect-size estimates (instead of simply focusing on the most significant variant for each associated locus) [52] and, more recently, combining association evidence from several traits to refine trait prediction [53,54]. Strategies integrating GWAS with gene expression data, such as **transcriptome-wide**

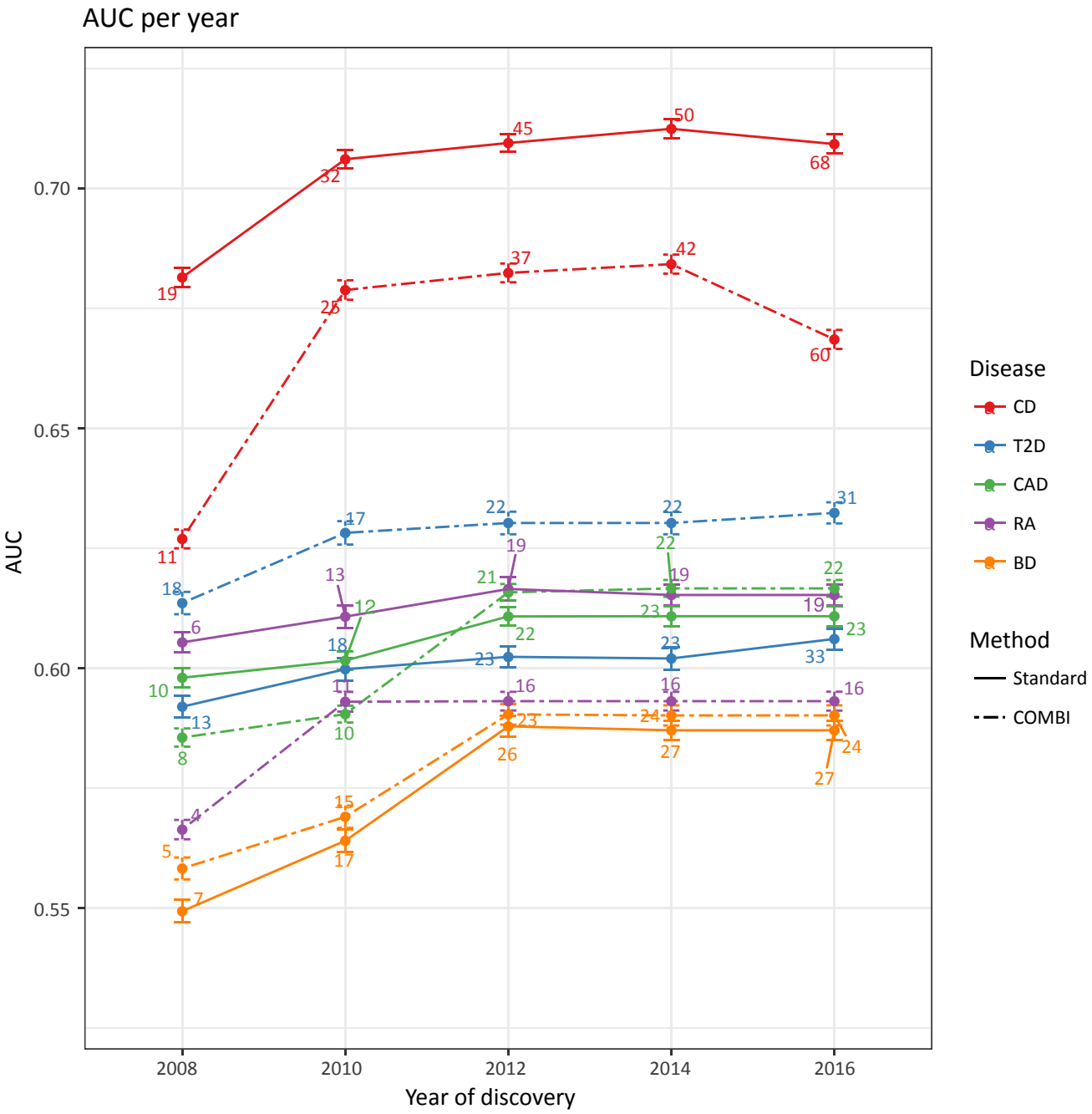


Figure 3. Increased Availability of Known Risk Variants and New Methodologies Are Improving the Ability to Classify Individuals According to Disease. We calculated the area under the ROC curve (AUC) for five diseases using the samples from the Wellcome Trust Case Control Consortium (WTCCC) paper and SNPs discovered in different year ranges. X-axis: Discovery year of disease-associated variants included in the genetic risk score (GRS) for each disease (based on year of inclusion in the GWAS Catalog). Y-axis: AUC using samples from the WTCCC case–control study. Performance of two predictors is shown; namely, a standard GRS (broken lines) and a modified version of the GRS using the number of risk alleles multiplied by the weight for each SNP, as derived from the COMBI algorithm (unbroken lines). The numbers of SNPs used in each AUC calculation are indicated in the small numbers next to each point. Even if the number of risk SNPs has increased steadily, for both methodologies we observe a plateau in the predictive power of the GRSs.

association studies (TWAS) [55], or directly based on the latter such as the transcriptional risk score (TRS) [56] also hold great promise for increasing the prediction of phenotypes from genotypes.

Still, it is important to ensure that the community maintains realistic hopes about the potential of genetic predictors. Ultimately, even if all causal variants and their true effects were known with exactitude, the heritability of each disease constitutes an upper bound for the maximum phenotypic variance that could be accounted for by a genetic predictor [50]. For instance, PRSs for T2D will top off with AUC ~0.8 as GWAS progressively grow into sample sizes in the low millions [57].

More importantly, the foreseeable improvements in terms of AUC do not imply immediate translational potential in the clinical setting. Most complex diseases have a low prevalence, and therefore even a specific cut-off value in risk score that maximizes both the detection of future

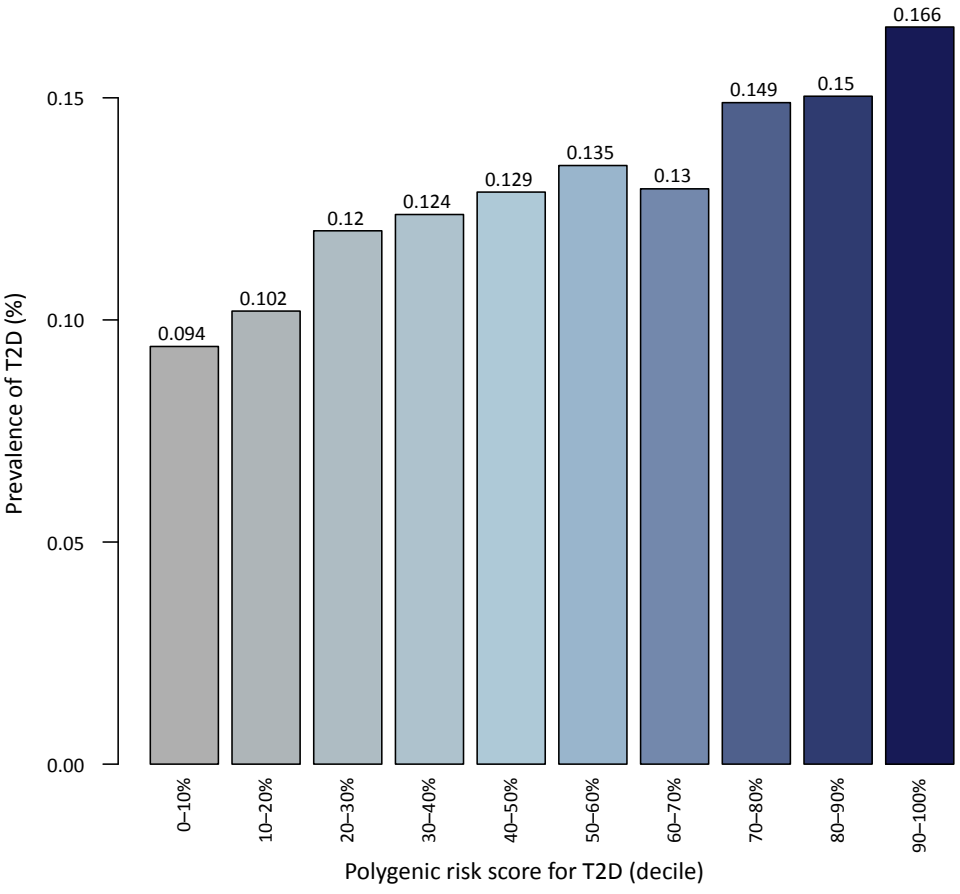


Figure 4. Polygenic Risk Scores (PRSs) Stratify Individuals According to Risk of Disease. X-axis: 55 210 samples of European ancestry from the Kaiser GERA Cohort are classified according to deciles of PRS for type 2 diabetes (T2D). Y-axis: Percentage of patients in each category. The graph shows the impact of increasing deciles of a weighted genetic risk score based on 414 linkage disequilibrium (LD)-pruned SNPs associated with T2D ($P < 10^{-3}$ in the 2014 transethnic DIAGRAM GWAS). The PRS captures risk of disease according to the genetic makeup of individuals, with twofold enrichment of cases in the top versus the lowest decile.

cases and the exclusion of healthy individuals can be offset by the fact that a large proportion of the detected individuals end up being false positives. The latter point is summarized by the positive predictive value (PPV); namely, the proportion of ascertained individuals that will truly develop the disease, which can remain low even if the AUC increases. A low PPV hurts the diagnostic power of risk scores because prediction of disease is not accurate at the individual level and thus neutralizes some of the social and economic benefits of genetic testing for complex disease.

At present, the most immediate clinical application of risk scores lies in the potential for risk stratification and complements the information provided by other risk factors. To illustrate this point, we calculated a PRS for T2D in the Kaiser RPGEH GERA cohort. As shown in Figure 4, the classification of individuals according to PRS deciles effectively captures an increasing fraction of T2D cases, which hints at its potential for stratification according to individual risk of disease. It is often underappreciated that many epidemiological risk factors have effect sizes that are like those of risk alleles discovered by GWAS, such as dietary factors that are known to increase risk of cancer [58]. This suggests that, rather than genomics information substituting for classical epidemiology as the initial fad seemed to indicate, we are entering an era of integrative predictive medicine. In conclusion, while it is still too soon, we can anticipate that genetic profiling will help in pinpointing individuals at high risk, which overall should lead to targeted lifestyle interventions and better decision-making in the clinic.

Concluding Remarks and Future Perspectives

GWAS have profoundly revolutionized both medical and complex trait genomics. To a large extent, this has been possible because findings in these fields, for the first time, are trustworthy. Although a variety of reasons account for this success, it is important to highlight the beneficial effect of having access to a robust methodology. GWAS led to the genetics community taking measures to avoid practices that lead to false discoveries; for example, by enforcing strict *P*-value thresholds corrected for multiple testing to declare findings statistically significant. Although we have witnessed amazing discoveries, we need new visions and methodologies to fully tackle questions about the genetic architecture of complex traits. In particular, larger studies and improved methods will be needed to keep improving phenotype prediction to the level where it is justifiably commonplace in clinical settings. How quickly this will be fully achieved, rather than whether it will or not, is one of many unanswered questions (see Outstanding Questions).

Outstanding Questions

Will the field grow tired of GWAS in response to new paradigms and, if so, will this decrease the interest in replicability that the field has seen in the past decade?

If, instead, GWAS are scaled up to sample sizes in the order of millions, how will replicability be ensured in face of the increased levels of phenotypic and genetic heterogeneity? What will be the accuracy of genetic risk predictors based on these large-scale studies?

What will be the replicability of findings arising from new methodologies such as machine learning or studies inspecting complex phenomena such as gene-by-environment and epistatic interactions? The extent to which good practices that fueled GWAS success can be implemented on these studies is unknown.

In what form will GWAS findings from genomic medicine be incorporated in the day-to-day clinical setting? The extent of the accuracy they will bring to practice remains open.

Regarding other scientific fields, whether a generalized push for larger sample sizes will be enough to ensure replicability remains to be seen. More complex measures such as registering studies before completion (as is done in clinical trials) should also help.

Acknowledgments

A.N. and J.A.R. were supported by the Ministerio de Ciencia e Innovación, Spain (BFU2015-68649-P, MINECO/FEDER, UE), the Direcció General de Recerca, Generalitat de Catalunya (2014SGR1311 and 2014SGR866), the Spanish National Institute of Bioinformatics (PT13/0001/0026), and the REEM (RD16/0015/0017) of the Instituto de Salud Carlos III, grant MDM-2014-0370 through the "María de Maeztu" Programme for Units of Excellence in R&D to UPF's Department of Experimental and Health Sciences). The authors have also received funding from the EU's Horizon 2020 research and innovation program 2014-2020 under Grant Agreement No. 634143 (*MedBioinformatics*). U.M.M. and G.G. were funded by US NIH grants 1-P01-GM099568 (Project 3) and 2-R01-DK087694.

Resources

[i]www.ebi.ac.uk/gwas/

[ii]www.ebi.ac.uk/gwas/diagram

Supplemental Information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tig.2018.03.005.

References

1. Visscher, P.M. *et al.* (2017) 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22
2. Maher, B. (2008) Personal genomes: the case of the missing heritability. *Nature* 456, 18–21
3. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678
4. Boyle, E.A. *et al.* (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186
5. Ioannidis, J.P. (2005) Why most published research findings are false. *PLoS Med.* 2, e124
6. Hilgard, J. and Jamieson, K.H. *et al.* (2017) Science as "broken" versus science as "self-correcting": how retractions and peer-review problems are exploited to attack science. In *The Oxford Handbook of the Science of Science Communication* (Jamieson, K.H., ed.), pp. 85–92, Oxford University Press
7. Plomin, R. *et al.* (2016) Top 10 replicated findings from behavioral genetics. *Perspect. Psychol. Sci.* 11, 3–23
8. Kraft, P. *et al.* (2009) Replication in genome-wide association studies. *Stat. Sci.* 24, 561–573
9. Visscher, P.M. *et al.* (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24
10. Marigorta, U.M. and Navarro, A. (2013) High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* 9, e1003566
11. Palmer, C. and Pe'er, I. (2017) Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet.* 13, e1006916
12. Waters, K.M. *et al.* (2009) Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. *Cancer Epidemiol. Biomarkers Prev.* 18, 1285–1289
13. Waters, K.M. *et al.* (2010) Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genet.* 6, e1001078
14. Carlson, C.S. *et al.* (2013) Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* 11, e1001661
15. Fu, J. *et al.* (2011) Multi-ethnic studies in complex traits. *Hum. Mol. Genet.* 20, R206–R213
16. de Candia, T.R. *et al.* (2013) Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. *Am. J. Hum. Genet.* 93, 463–470
17. Manolio, T.A. (2017) In retrospect: a decade of shared genomic associations. *Nature* 546, 360–361
18. Martin, A.R. *et al.* (2017) Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* 100, 635–649
19. Nakagome, S. *et al.* (2012) Crohn's disease risk alleles on the *NOD2* locus have been maintained by natural selection on standing variation. *Mol. Biol. Evol.* 29, 1569–1585
20. Wang, Y.F. *et al.* (2007) Clinical manifestations of inflammatory bowel disease: East and West differences. *J. Dig. Dis.* 8, 121–127
21. CONVERGE Consortium (2015) Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* 523, 588–591
22. Klein, R.J. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385–389
23. Fritsche, L.G. *et al.* (2016) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* 48, 134–143
24. Cotsapas, C. *et al.* (2011) Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* 7, e1002254
25. Moser, G. *et al.* (2015) Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* 11, e1004969
26. Luciano, M. *et al.* (2018) Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* 50, 6–11
27. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research *et al.* (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316, 1331–1336
28. Scott, L.J. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316, 1341–1345
29. Timpson, N.J. *et al.* (2009) Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. *Diabetes* 58, 505–510
30. Kulminski, A.M. *et al.* (2016) Explicating heterogeneity of complex traits has strong potential for improving GWAS efficiency. *Sci. Rep.* 6, 35390
31. Magosi, L.E. *et al.* (2017) Identifying systematic heterogeneity patterns in genetic association meta-analysis studies. *PLoS Genet.* 13, e1006755
32. Hirschhorn, J.N. *et al.* (2002) A comprehensive review of genetic association studies. *Genet. Med.* 4, 45–61
33. Lohmueller, K.E. *et al.* (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* 33, 177–182
34. Ioannidis, J.P. *et al.* (2011) The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology* 22, 450–456
35. Siontis, K.C. *et al.* (2010) Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *Eur. J. Hum. Genet.* 18, 832–837
36. Ioannidis, J.P. *et al.* (2006) Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am. J. Epidemiol.* 164, 609–614
37. Weiss, K.M. and Terwilliger, J.D. (2000) How many diseases does it take to map a gene with SNPs? *Nat. Genet.* 26, 151–157
38. Contreras, J.L. (2015) NIH's genomic data sharing policy: timing and tradeoffs. *Trends Genet.* 31, 55–57
39. Anon (2012) Asking for more. *Nat. Genet.* 44, 733
40. Chanock, S.J. *et al.* (2007) Replicating genotype–phenotype associations. *Nature* 447, 655–660
41. Yang, J. *et al.* (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47, 1114–1120
42. Fadista, J. *et al.* (2016) The (in)famous GWAS *P*-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* 24, 1202–1205
43. Keller, M.C. (2014) Gene × environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol. Psychiatry* 75, 18–24
44. Marigorta, U.M. and Gibson, G. (2014) A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. *Front. Genet.* 5, 225
45. Mieth, B. *et al.* (2016) Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Sci. Rep.* 6, 36671
46. Pickrell, J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* 94, 559–573
47. Wood, A.R. *et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46, 1173–1186
48. Rodriguez, J.A. *et al.* (2014) Integrating genomics into evolutionary medicine. *Curr. Opin. Genet. Dev.* 29, 97–102
49. So, H.C. and Sham, P.C. (2017) Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits. *Bioinformatics* 33, 886–892
50. Wray, N.R. *et al.* (2013) Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* 14, 507–515

51. Robinson, M.R. *et al.* (2017) Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* 1, 0016
52. Vilhjalmsson, B.J. *et al.* (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592
53. Turley, P. *et al.* (2018) Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* 50, 229–237
54. Maier, R.M. *et al.* (2018) Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat. Commun.* 9, 989
55. Gusev, A. *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252
56. Marigorta, U.M. *et al.* (2017) Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat. Genet.* 49, 1517–1521
57. Chatterjee, N. *et al.* (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* 45, 400–405 405e1–405e3
58. Figueiredo, J.C. *et al.* (2014) Genome-wide diet–gene interaction analyses for risk of colorectal cancer. *PLoS Genet.* 10, e1004228
59. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517
60. Carney, D.R. *et al.* (2010) Power posing: brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychol. Sci.* 21, 1363–1368
61. Ranehill, E. *et al.* (2015) Assessing the robustness of power posing: no effect on hormones and risk tolerance in a large sample of men and women. *Psychol. Sci.* 26, 653–656
62. Rousseau, D.L. (1971) "Polywater" and sweat: similarities between the infrared spectra. *Science* 171, 170–172
63. Caspi, A. *et al.* (2003) Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* 301, 386–389
64. Szucs, D. and Ioannidis, J.P. (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* 15, e2000797
65. Benjamin, D.J. *et al.* (2018) Redefine statistical significance. *Nat. Hum. Behav.* 2, 6–10
66. Patil, P. *et al.* (2016) What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* 11, 539–544
67. Jager, L.R. and Leek, J.T. (2014) An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 15, 1–12